利用 N-gram 和语义分析的维吾尔语文本相似性检测方法 *

张 莹 [1]，亚森·艾则孜 [1†]，吴顺祥 [2]

(1. 新疆警察学院 信息安全工程系，乌鲁木齐 830013; 2. 厦门大学 自动化系，福建 厦门 361005)

摘 要：目前自然语言文本相似度估计大多是针对英语等一些大类语言，为了实现维吾尔语文本的相似性检测，提出一种基于 N-gram 和语义分析的相似性检测方法。首先，根据维吾尔语单词特征，采用了 N-gram 统计模型来获得词语，并根据词语在文本中的出现频率来构建词语-文本关系矩阵，作为文本模型。然后，采用了潜在语义分析(LSA)来获得词语及其文本之间的隐藏关联，以此解决维吾尔语词义模糊的问题，并获得准确的相似度。在包含重组和同义词替换的剽窃文本集上进行实验，结果表明该方法能够准确有效地检测出相似性。

关键词：维吾尔语；文本相似性检测；N-gram 统计模型；潜在语义分析

中图分类号：TP391 doi: 10.3969/j.issn.1001-3695.2018.03.0158

Uyghur text similarity detection method using N-gram and semantic analysis

Zhang Ying[1], Yasen Aizezi[1†], Wu Shunxiang[2]

(1. *Dept. of Information Security Engineering Xinjiang Police College*, *Urumqi Xinjiang* 830013, *China*; 2. *Dept. of Automation*, *Xiamen University*, *Xiamen Fujian* 361005, *China*)

Abstract: At present, most of the researches on the similarity of natural language texts are aimed at some major languages such as English. In order to detect similarities between Uighur texts, this paper proposed a similarity detection method based on N-gram and semantic analysis. Firstly, it used N-gram statistical model to obtain the words based on Uyghur word features, and constructed the word-text relation matrix according to the appearance frequency of the words in the text. Then, it adopted a latent semantic analysis (LSA) to obtain the hidden association between the words and their texts, so as to solve the problem of vague semantic meaning in Uyghur language and obtain exact similarity. Experiments on plagiarized text sets containing reorganization and synonym replacement show that this method can detect the similarity accurately and effectively.

Key words: Uyghur language; text similarity detection; N-gram statistical model; latent semantic analysis

0 引言

通过网络获取信息十分容易，这使学术剽窃成为一种简单操作。存在以下几种类型的文档剽窃[1]：a)直接从发表的文本中复制短语或段落，而不给出引用出处和作者；b)将已发表内容进行语句和结构修改并进行使用。为了保护作者的版权，对待发表文档进行剽窃检测是一种重要手段[2]。一些相似度估计和剽窃检测方法是与语言无关的，可以适用于多种语言，而另一些则是对语言比较敏感。语言无关的方法是基于文本特征的估计，这些特征不是特定自然语言固有的，如单个字符的数量和平均句子长度值等[3]。而语言敏感的方法是基于单一语言特定属性的，比语言无关的方法具有更高的针对性和准确性。

近些年，随着新疆经济和教育的发展，产生了很多以维吾尔语进行书写的学术论文[4]。对维语文档进行相似度计算和剽窃检测对维语文化的健康发展具有重要意义。本文是针对维吾尔语文本相似性的检测，由于维吾尔语词语可能有多种词形变化、同义词和不同含义。比如，每个词有不同形态，前缀和后缀可以以连续的方式附加到单词上。单个字符串可能包含动词变形、介词变形、代词变形和连词变形等。因此，维吾尔语文本单词的语义比较模糊，给剽窃检测造成了一定的难度[5]。

由于维吾尔语文档的信息化处理发展较晚，目前在维吾尔语文档相似性等方面的研究单位主要为新疆大学。由于维吾尔语的复杂语言结构，一些常用的相似性度量都不能很好地应用[6]。目前很少有学者提出相关方法，其中文献[7]提出一种维吾尔语句子相似度计算方法(MUSM)，其采用词形特征，通过多策略精选算法来计算两个维吾尔语句子的相似度。然而，其只能

收稿日期：2018-03-12；**修回日期：**2018-05-07 **基金项目：**国家自然科学基金资助项目（61762086）；新疆维吾尔自治区高校科研计划立项项目（XJEDU2016S090）

作者简介：张莹（1988-），女，山东梁山人，讲师，硕士，主要研究方向为偏微分方程及应用、计算机应用；亚森·艾则孜（1975-），男（维吾尔族，通信作者），新疆库车人，国家电子数据司法鉴定员，教授，硕士，主要研究方向为数字取证、自然语言处理等；吴顺祥（1967-），男，湖南邵阳人，国家电子数据司法鉴定员，教授，博士，主要研究方向为智能信息处理与信息内容安全.

在句子级进行检测，且没有考虑到同义词的替换问题。文献[8]首先引入和分析了维吾尔语文本语义相似性度量，通过上下文来确定语义相似度，可以应对同义词的问题，但是其精确性较低。

本文提出一种基于 N-gram 和语义分析的维吾尔语文本相似性检测方法。其主要创新点为：a)根据维吾尔语单词特征，采用了 N-gram 统计模型来获得词干，并根据单词频率来构建文本模型；b)为了解决维吾尔语单词词义模糊的问题，采用了潜在语义分析(latent semantic analysis，LSA)来获得词语及其文本之间的隐藏关联，获得相似度。实验结果表明，提出的方法能够准确有效地检测出包含重组和同义词替换的剽窃文本。

1 提出方案的基本框架

1.1 维吾尔语特征

维吾尔语是以阿拉伯字母为基础的文字，具有高度的黏着性。维吾尔字母共有 32 个，字母的形式具有多样性，通常包含 4 种表现形式，致使其形态变化较为复杂。维吾尔语单词由词干和词缀组成，在同一词干前后添加不同的词缀可以表示不同的词义[9]。由于这些特征，给维吾尔语文本信息处理造成一定的困难，如特征维数大[10]。

表 1 展示了在词干“ئەدىب”（作者）的前后添加不同词缀所形成的词语及其含义，其中，下划线划出的为词缀。

表 1 词干“ئەدىب”(作者)上添加词缀形成的词语

词语	词义	词语	词义
ئەدىب	作者	ئەدىبنىڭ	作者的
ئەدىبە	作者（女）	ئەدىبەنىڭ	作者的（女）
ئەدىبلەر	作者们	ئەدىبتەك	像那个作者
ئەدىبەلەر	作者们	ئەدىبىم	我的作者

1.2 基本框架

本文目标是开发一种用于自然语言文本的相似度分析方法。提出的方法可以采取两种工作模式。第一种模式，分析文本之间的相似度，包括可疑和参考文本；而第二种模式包含一个输入，即为基于文本的查询，输出是文本或查询之间的相似度度量。

提出的方法中，假设原始文本和重写文本都具有可衡量的差异，这些差异可以通过统计和语言指示器来获取。为了克服维吾尔语文本中的相似度/剽窃检测的困难，本文主要采用了三个技术手段。第一个是采用了自然语言处理（natural language processing，NLP）技术，而不是依赖于传统的字符串匹配方法。第二个是采用了能够克服大量词汇和句法挑战的文本建模技术。第三个是使用了潜在语义分析(LSA)来确定文本中包含的隐藏关联。其中，为了能从给定的文本中推断出潜在语义，进行大量的统计计算，本文考虑了奇异值分解（singular value decomposition，SVD）。提出的文本相似度分析方法的整体步骤如图 1 所示。

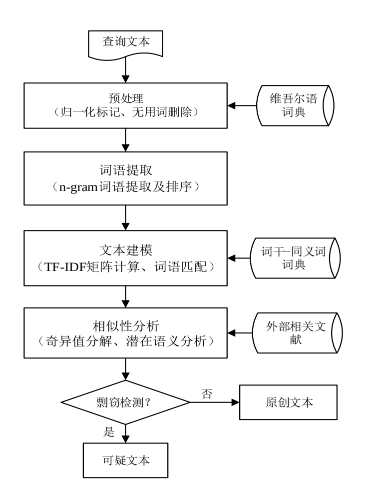


图 1 提出的文本相似度分析方法的框架

提出的方法主要由以下几个部分组成，包括文本预处理、词语提取、文本建模和相似性分析。其中，预处理阶段包括文本归一化标记、无用词删除等操作。词语提取阶段主要包括利用 N-gram 技术的词语提取和排序。文本建模阶段包括文本的 TF-IDF 矩阵计算和词语匹配。相似度阶段包括奇异值分解和潜在语义分析。

预处理过程中，索引模块逐个读取文本，为每个语句生成单词索引，并将这些索引传递给 N-gram 计数模块。N-gram 计数模块将每个文本生成的 N-gram 词语写入单独的临时文件。这些临时文件被合并到一个文件中，并且对 N-gram 结构进行排序，去除重复计数。接下来，文本建模模块读取排序的 N-gram 结构文件以计算 TF-IDF 矩阵，该矩阵作为给定文本集的特征矩阵。然后，相似度估计模块通过特征矩阵计算文本间的余弦相似度，作为文本相似度的初步估计。接着，将 TF-IDF 矩阵传递给 LSA 函数，实现对文本集相似度进行深度估计。主要数据处理流程图如图 2 所示。

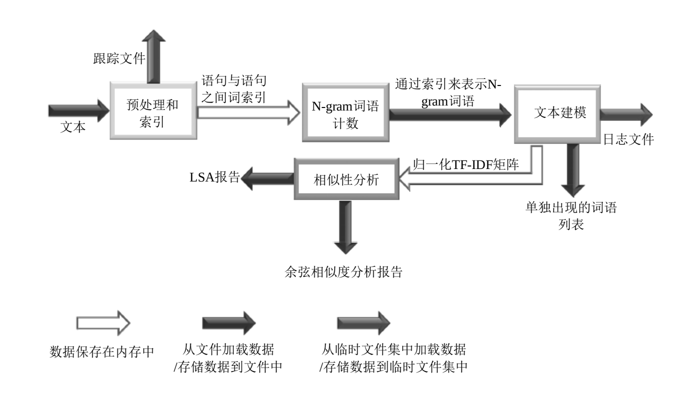


图 2 相似度估计的数据处理流程图

2 文本相似度估计步骤

2.1 文本预处理

文本预处理是自然语言处理任务成功实现的重要前提。首先，将输入文本集转换为纯文本，文本中所有控制字符需进行过滤。接下来，解析文本以进行 PoS 标记，在这项工作中使用了维吾尔语言模型。

对于每个文本，标记每个语句并且将其存储在存储器中，文本标记案例如图 3 所示维吾尔文文本“بۇ سۇ بەك ياخشى”（译：这口水真好）。通过调用形态分析器来获取词语索引，以获得每个变形词的同类分析。这些分析是用来消除歧义的，对每个变形词采用相关的词性（PoS）标签。应用 PoS 标注可以解决可能存在的单词形态模糊性。如果仍然有多个可能的同类词，则使用 Levenshtein 编辑距离[11]用于选择最可能的词干，其变形词与可能的词干之间的编辑距离最小。在这项工作中，使用了文献[12]采用的形态分析器以及维吾尔语词法查询。这种形态分析器是基于语言学方法开发的，根据所选择的词干，使用存储在词典中的词干索引来对变形词进行索引。

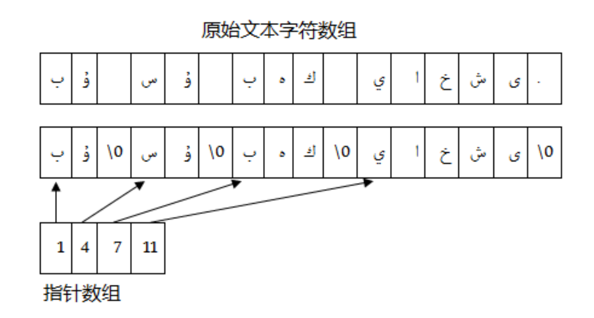


图 3　文本标记例子

索引时删除停止词，这是通过检查每个变形词的形态特征来实现的。如果这些特征的值表示当前的词是感叹词、介词或代词，则认为变形词是停止词。

2.2　基于 N-gram 统计模型的词语提取

在文本分类、检测等应用中，通常需要首先提取文本中的词语。本文采用更适合维吾尔语环境的统计方法来提取词语。所采用的统计方法为 N-gram 统计模型[13]。在字母层上进行单词切分，即将连续 N 个字母作为一个 gram 单元。

N-gram 模型中，对于文本中一个特定字母 l_i，设定其出现的概率与前面 N-1 个字母的出现情况相关。因此，字母序列 $L = l_1 l_2 l_3,\ldots,l_N$ 出现的概率为

$$P(L) = P(l_1 l_2 l_3,\ldots,l_N) = \prod_{i=1}^{N} P(l_i \mid l_{i-N+1},\ldots,l_{i-1}) \quad (1)$$

N-gram 模型中 N 的设定需要结合具体的语言环境，对于维吾尔语，由于其每个单词都由多个字母结合而成，为此较小的 N 不能有效地代表单词属性，而 N 较大如等于 3 或 4 时，则具有较强的代表性。

本文利用 N-gram 统计模型提取词语过程中，为了降低单词维度和冗余度，首先根据维吾尔语词典，删除了单词中最常见的词缀。然后，计算两个词语的相似度，以此来提取词干。

为了展示 N-gram 统计模型提取词语的过程，列举了一个 N=2 时的例子，即计算两个词ئىنقلاۋى（革命）和ئىنقلاۋلىغى（革命的）的相似度。

ئىنقلاۋلىغى ⇒ ئىن، قى، لا ، ۋى، لى، غى。（首先将词分解为 N=2 字母组合单元）

去除常用词缀的两字母组合 ⇒ ئىن، قى، لا ، ۋى 。

.ئىنقلاۋى ⇒ ئىن، قى، لا ، ۋى。

去除常用词缀的两字母组合 ⇒ئىن، قى، لا ، ۋى 。

那么，这两个单词的相似性为 $S = \dfrac{2C}{A+B} = \dfrac{2\times 3}{4+3} = 0.8571$。其中，$A$ 表示第一个单词中所包含的且第二个单词中不存在的字母组合的数量；同样，B 第二个单词中所包含的且第一个单词中不存在的字母组合的数量；C 表示两个词中都包含的相同字母组合的数量。若两个单词的相似性大于设定的阈值，则将这两个词合并为一个词干。

从预处理文本中提取指定长度的 N-gram 单词。最近的实验表明，N-gram 最合适的长度在 2~7[14]。对于所考虑的每个 N-gram 大小，词语提取的过程必须是连续的，例如，对于 unigram、bigram 和 trigram 程序必须运行三次。为了避免巨大的存储要求，从单个数据块中提取 N-gram 并且一次只将一个 N-gram 大小保存在存储器中，以使效率最大化。 N-gram 计数步骤的基本过程如图 4 所示。

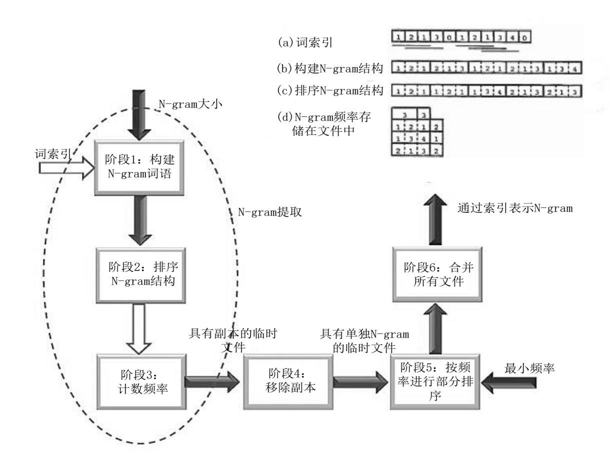


图 4 N-gram 计数的数据处理流程图

剽窃检测最有效的方法是将文本频率（DF）作为特征。为了减少文本中的词语数量，本文根据文本频率来确定一些词语是否重要。只在一个文本中存在的词语将被立即删除，因为它们不会在任何其他文本中被剽窃。此外，本文删除了包含在超过 $u+\sigma$ 个文本中的一些词语（其中 u 是平均文本频率，σ 是平均文本频率的标准偏差）。换句话说，即从文本中删除了所有常见的词语。

2.3　构建文本模型

本文考虑了文本中词语的出现频率，提出了一种词语文本模型。这些词语与文本的关系以矩阵形式表示，其中列表示文本，行表示词语。考虑由 m 个向量组成 $n\times m$ 矩阵 A $[A_1, A_2,\ldots,A_m]$，其中向量 A_j 表示包含在文本 j 中的词语。

每个向量 A_j 由 n 个元素 a_{ij} 组成，a_{ij} 表示文本 j 中词语 i 出现频率的权值，如式（2）所示。这个等式是本文提出的用于构造特征矩阵 A 的权值系数，是标准 TF-IDF 加权的修改版本。

$$a_{ij}=\begin{cases}\frac{1}{2}+\dfrac{PF_{ij}\cdot\log(\frac{|N|}{DF_i})}{2\cdot\max_j(PF_{ij})\cdot\log(|M|)}, & \text{if } i\in j\\ 0, & \text{others}\end{cases}\qquad(2)$$

其中：PF_{ij} 表示文本 j 中词语 i 出现的频率，DF_i 表示出现词语 i 的文本数量，$|M|$ 是所有文本的数量。与 TF-IDF 相比，提出的频率权重计算方法的差异在于 IDF 的标准化。本文将其除以 $\log(|M|)$，以使 $a_{ij}\in<0.5,1>$。另一方面，如果词语 i 不在文本 j 中出现，则 $a_{ij}=0$。这种加权机制有助于后续采用的 SVD 产生最佳效果。

在构建 TF-IDF 矩阵 A 期间，执行成对的 N-gram 词语匹配。这是一个直接比较的过程，其整体复杂度为 $O(N)$，其中 N 是所考虑文本中单独词语的数量。当考虑词汇和句法变化时，整个配对过程的复杂度将增加到 $O(N^2)$。对于这种情况，可以使用一些技术来估计成对词语匹配得分。在这项工作中，本文在这种匹配过程中使用了匹配平均和骰子系数。这是通过以矩阵形式表示每对标记词语之间的关系来进行的，以此来计算匹配得分。

表示两个词语成对匹配的矩阵“cost”的计算方法为：如果第一个词语内标记 i 的索引等于第二个词语及其同义词、反义词中标记 j 的索引，则 $\text{cost}_{ij}=1$；否则，$\text{cost}_{ij}=0$。匹配得分的值表示所考虑的两个词语是否等同。在这项工作中，如果匹配得分等于 1.00，则认为该对词语是等同的。

2.4　潜在语义分析

这个阶段用于推断出文本中包含的词语之间的潜在语义关联。LSA[15]是一种智能文本比较技术，使用数学算法分析大量文本，并揭示文本的底层语义信息，使其成为自然语言文本剽窃检测的可行技术。

本文使用一种将对称矩阵对角化的线性代数技术：奇异值分解（SVD），将矩阵 A 分解成三个独立的矩阵即左奇异矩阵 U、奇异矩阵 Σ 和右奇异矩阵 V。其中，矩阵 Σ 仅包含对角元素，称为奇异值，矩阵 U 和 V 包含分解的详细信息。

所有这些矩阵都可以在潜在空间 k 中被分解，以执行 A 的最佳 k 级近似，使得奇异值 $\sigma_{k+1},\sigma_{k+2},\ldots,\sigma_m$ 被替换为 0，其中 $1\le k\le m$。那么，矩阵 U 是 $n\times k$ 列正交矩阵，其列是词语奇异向量。Σ 是 $k\times k$ 对角矩阵，不包含表示奇异值的负数和零。

$$\sigma_1\ge\sigma_2\ge\cdots\ge\sigma_k\ge\sigma_{k+1}=\cdots=\sigma_m=0\qquad(3)$$

SVD 的一个特征是 Σ 对角线上的奇异值按降序排列，满足式（3）。矩阵 V^T 是一个 $k\times m$ 正交矩阵，其行是文本奇异向量。

图 5 呈现了 SVD 分解过程。在分解之后获得的矩阵 V^T 是进一步处理的基本构件，因为它包含文本的独立轮廓向量。

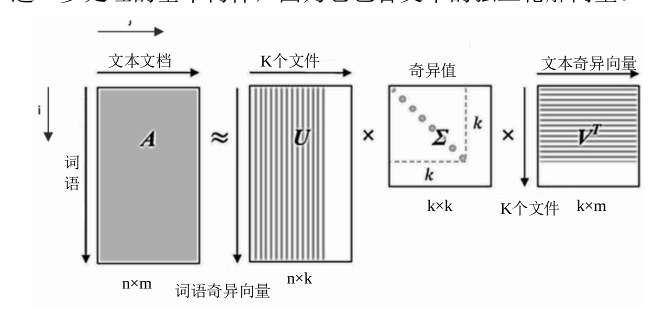


图 5 文本词语矩阵的奇异值分解

2.5　相似度估计

最后一个阶段为计算成对文本的相似度。在我们使用矩阵 V^T 之前，必须用相应的奇异值重新缩放所有文本配置文件的单个元素，如下所示。

$$B=\Sigma\times V^T\qquad(4)$$

然后，根据式（5）计算相关性。其中，矩阵 B 的列长度被标准化。所得到的 sim_{SVD} 是对称矩阵，其中每对文本由一个得分来表示相似度。

$$\text{sim}_{SVD}=\|B\|^T\times\|B\|\qquad(5)$$

减少词语有助于文本得到更高的 sin_{SVD} 得分，其中绝大多数短语是无意义的，需要被删除。因此，本文修改矩阵 sin_{SVD} 的计算式，如式（6）所示。所得到的估计结果为对应相似度测量的总体情况。

$$\text{sim}(R,S)=\text{sim}_{SVD}(R,S)\cdot\frac{\sqrt{|N_{red}(R)|\cdot|N_{red}(S)|}}{\min(|N_{orig}(R)|\cdot|N_{orig}(S)|)}<\tau\qquad(6)$$

其中，τ 是相似度阈值。

如果使用与查询中指定类型相对应的加权频率来计算查询向量 q，则其表示一个与矩阵 A_k（与原始 $n\times m$ 词语文本矩阵 A 相对应）的列相比较的可疑文本。假设向量 e_j 表示维度为 m 的第 j 个规范向量（即，$m\times m$ 单位矩阵 I_m 的第 j 列）。因此，向量 $A_k e_j$ 是秩为 k 的矩阵 A_k 的第 j 列。对于文本向量 $b_j=\sum_k V_k{}^T e_j$，查询向量 q 和 A_k 的 m 维文本向量（或列）之间的夹角余弦值可以由以下公式表示：

$$\cos\theta_j=\frac{b_j{}^T(U_k{}^T q)}{\|b_j\|_2\|U_k{}^T q\|_2},\quad j=1,2,\ldots,m\qquad(7)$$

可以通过设置 A 中所有除了 k 个最大值以外的奇异值等于零，来近似构造 A 的秩 k，其中 $k\le r_A$ 并且 r_A 为 Σ 中非零对角线元素的数量。A_k 与 A 的近似误差为

$$\|A-A_k\|_F=\min_{rank(B)\le k}\|A-B\|_F=\sqrt{\sigma_{k+1}^2+\cdots\sigma_{r_A}^2}\qquad(8)$$

其中，$A_k = U_k \sum_k V_k^T$，U_k 和 V_k 分别是 U 和 V 的第 k 列，$\sum_k$ 是 $k \times k$ 对角矩阵，包含 A 中 k 个最大奇异值。换句话说，原始词语文本矩阵 A 与 A_k 的近似误差由截断的（或丢弃的）奇异值（$\sigma_{k+1}, \sigma_{k+2}, \ldots, \sigma_{r_A}$）确定。通过 A_k 近似 A 所反映的相对变化由 $\frac{\|A - A_k\|_F}{\|A\|_F}$ 估计，其中，实数 $m \times n$ 矩阵 $B = [b_{ij}]$ 的 Frobenius 矩阵范数（$\|\cdot\|_F$）被定义为：$\|B\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} b_{ij}^2}$。

3　实验及分析

3.1　实验设置

为了估计所提出的方法在估计维吾尔语文本与潜在文字剽窃（包括词语重组和同义词替换）文本之间相似度的性能，使用包含 9 个维吾尔语文本(L1~L9)的数据集来进行测试。文本中包含了特定数量的无用停止词与有用词语，如图 6 所示。

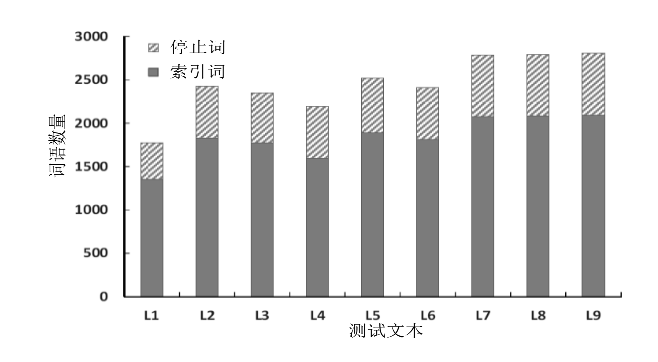


图 6　用于相似度估计的 9 个文本

为了构建包含重新组合和同义词替换的剽窃文本，本文设定的数据集中前 5 个文本为原件文本，第 6 个文本是从第 3 个文本中提取的。第 7 个文本由两部分组成，一部分来自于第 3 个文本，另一部分来自于第 4 个文本。第 8 个文本是第 7 个文本的精确副本，但是 50％的词被更改为其同义词。最后一个文本是从第 7 个文本中生成的，但对 50％的语句进行了重组。图 7 显示了所考虑的 9 个文本的实际相似度关系。

	L1.txt	L2.txt	L3.txt	L4.txt	L5.txt	L6.txt	L7.txt	L8.txt	L9.txt
L1.txt	N/A	3%	4%	6%	3%	4%	6%	6%	7%
L2.txt	2%	N/A	3%	3%	3%	3%	5%	5%	5%
L3.txt	2%	3%	N/A	3%	5%	100%	45%	45%	45%
L4.txt	5%	4%	3%	N/A	3%	3%	70%	70%	70%
L5.txt	2%	3%	5%	2%	N/A	5%	5%	4%	5%
L6.txt	2%	3%	100%	3%	5%	N/A	45%	45%	44%
L7.txt	4%	5%	47%	56%	5%	47%	N/A	100%	100%
L8.txt	4%	5%	48%	56%	5%	48%	100%	N/A	100%
L9.txt	5%	5%	47%	55%	5%	47%	100%	100%	N/A

图 7　9 个文本的实际相似度

所有实验在 Intel Core i7-4700 CPU，主频 2.4 GHz，微软 Windows 8 系统平台上，通过 MATLAB 编译实现本文算法，并进行相似度估计。

3.2　参数选择

为了获得 N-gram 算法中最优的 N 值，设定其值在 1 到 6 之间时，测量相似度估计的精确度、召回率和 F-measure 值。其中，设置剽窃文本相似度阈值 τ 为 30%，即当两个文本之间的相似度达到 30%时，即为剽窃。

为了取得统计意义上的比较结果，在 9 个文本上重复进行了 30 次实验，平均结果如图 8 所示。可以看出，不同 N 值下算法的检测性能不一样。当 N 值较大和较小时性能都不理想，但当 N=3 或 4 能够取得较为优越的结果。这是因为 N 较小时，获得的词不能足够表达真实含义。当 N 较大时，增加了近似矩阵 A 的语义维数，对相似度度量估计具有负面影响。

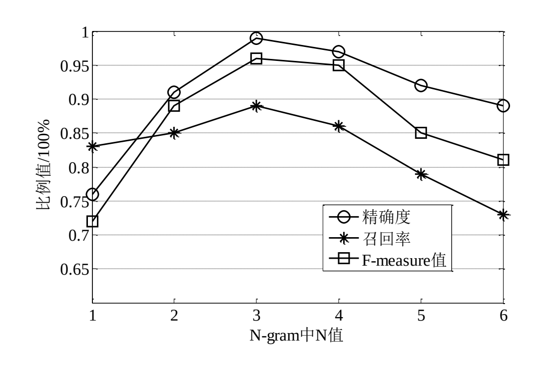


图 8　本文方法在不同 N 值下的性能指标

为了更加准确地展现性能差异，参考真实相似度值，统计本文方法 30 次实验所估计的成对文本相似度度与真实值的绝对差的最大值和平均值，如表 2 所示。可以得出结论：N-gram 中使用 N=3 获得的结果要优于其他结果。

表 2　相似度估计值与真实值的最大差值和平均差值

N-gram 值	最大|差|值	平均|差|值
N=1	28.74%	12.82%
N=2	17.66%	3.32%
N=3	12.27%	2.25%
N=4	13.62%	2.57%
N=5	21.72%	3.54%
N=6	25.87%	4.71%

另外，随着 N 值的增加，本文方法的计算时间也会增加，如图 9 所示。为此，在综合考虑检测性能和检测时间情况下，最终选择 N=3。

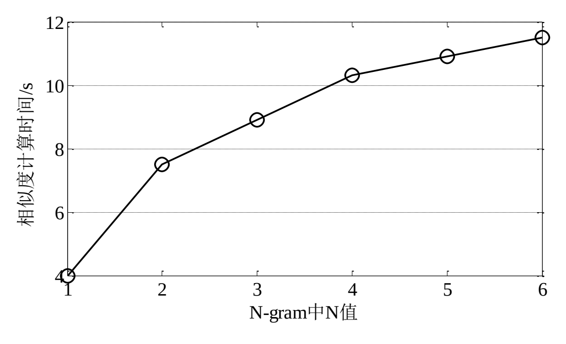


图 9　本文方法在不种 N 值下的计算时间

3.3　性能比较

将本文提出的检测方法与文献[8]提出的相似性检测方法进行比较，其中本文方法中设置 N=3。同样在上述 9 个文本中进行实验，相似度估计和检测性能结果如表 3 所示。

表 3　性能比较结果

方法		文献[8]方法	本文方法
相似度估计	|差|的最大值	16.63%	12.27%
	|差|的平均值	2.58%	2.25%
检测性能	精确度	92.8%	99.4%
	召回率	85.6%	88.6%
	F-measure 值	92.5%	95.5%

可以看出，本文方法优于文献[8]方法，这是因为提出的方法是基于自然语言处理，能够很好地适合维吾尔语这种复杂语言。而文献[8]方法在估计相似度时包括了停止词，使其相似度估计值较高。

通常，在不存在同义词替换的情况下，本文方法与文献[8]方法对相似量的估计结果相近，特别是对于重组的文本集。而在同义词替换情况下，对于所有合成剽窃文本进行相似度测量时，所提出的方法比文献[8]方法表现更好。

因此，综上实验结果表明，本文方法能够准确估计具有形态变化、重组和同义词替换的维吾尔语文本相似度。

4　结束语

本文提出了一种专门用于维吾尔语文本相似度估计的方法。基于文本与 N-gram 词语之间的关系模型，对检查文本使用 PoS 标记，以支持在文本归一化期间解决形态歧义问题。通过文本索引和停止词删除，以构建文本 TF-IDF 模型。最后，使用 LSA 和 SVD 研究文本与词语之间的隐藏关联。实验结果证明，提出的方法在检测文字相似度方面表现出很强的能力，能够应对复制、句子重排和同义词替换等剽窃。本文相似性检测方法是针对维吾尔语文本提出，为此采用了 N-gram 统计模型来获得词干，对于英语、汉语等文本则无需采用这种词干提取方法。另外，所采用的基于 LSA 获得词语及文本之间隐藏关联的方法可适用于其他语言本文，有助于提高对具有模糊性语言文本的检测性能。

在未来工作中，将尝试更有效的词语匹配方法，以提高在检测语句变化情况下的效率。另外，还将考虑采用并行算法来处理大规模的文本。

参考文献：

[1] 邹杜，陈育青，张凌．基于语义匹配的抄袭检测方法 [J]．华南理工大学学报：自然科学版, 2013, 41 (7): 131-136. (Zou Du, Chen Yuqing, Zhang Ling. A Plagiarism detection method based on semantic matching [J]. Journal of South China University of Technology: Natural Science Edition, 2013, 41 (7): 131-136.)

[2] 张超，陈利，李琼．一种 PST_LDA 中文文本相似度计算方法 [J]．计算机应用研究, 2016, 33 (2): 375-377. (Zhang Chao, Chen Li, Li Qiong. Chinese text similarity algorithm based on PST_LDA [J]. Application Research of Computers, 2016, 33 (2): 375-377.)

[3] Barron-Cedeno A, Gupta P, Rosso P. Methods for cross-language plagiarism detection [J]. Knowledge-Based Systems, 2013, 50 (1): 211-217.

[4] 吐尔地•托合提，维尼拉•木沙江，艾斯卡尔•艾木都拉．基于语义串抽取及主题相似度度量的维吾尔文文本分类 [J]．中文信息学报, 2017, 31 (4): 100-107. (Turdi Tohti, Winira Musajan, Askar Hamdulla. Semantic string-based topic similarity measuring approach for Uyghur text classification [J]. Journal of Chinese Information Processing, 2017, 31 (4): 100-107.)

[5] Sindhu L, Idicula Sumam Mary. A plagiarism detection system for malayalam text based documents with full and partial copy [J]. Procedia Technology, 2016, 25 (4): 372-377.

[6] 买买提依明•哈斯木，吾守尔•斯拉木，维尼拉•木沙江，等．基于 N 元模型的维吾尔语文本分类技术研究 [J]．计算机应用研究, 2015, 32 (7): 1986-1988. (Maimaitiyiming Hasimu, Wushouer Silamu, Weinila Mushajiang, *et al.* Research N-gram based Uyghur text classification technique [J]. Application Research of Computers, 2015, 32 (7): 1986-1988.)

[7] 田生伟，吐尔根•依布拉音，禹龙，等．一种维吾尔语句子相似度算法的研究 [J]．计算机工程与应用, 2009, 45 (26): 144-146. (Tian Shengwei, Turgun Ibrahim, Yu Long, *et al.* Similarity measure algorithm of Uyhur sentence [J]. Computer Engineering and Applications, 2009, 45 (26): 144-146.)

[8] Ma Bo, Zhou Xi, Yang Yating, *et al.* Uyghur semantic similarity computation based on contextual information in web documents [J]. Journal of Computational Information Systems, 2012, 8 (2): 563-570.

[9] Yan Chenggang, Xie Hongtao, Liu Shun, *et al.* Effective Uyghur language text detection in complex background images for traffic prompt identification [J]. IEEE Trans on Intelligent Transportation Systems, 2017, 19 (1): 1-10.

[10] Mi Chenggang, Yang Yating, Wang Lei, *et al.* Detection of loan words in Uyghur texts [J]. Communications in Computer & Information Science, 2014, 49 (6): 103-112.

[11] 蒋宗礼，王威．融合检索技术的译文推荐系统 [J]．哈尔滨工程大学学报，2017, 38 (3): 419-424. (Jiang Zongli, Wang Wei. Translation recommendation system with information retrieval technology [J]. Journal of Harbin Engineering University, 2017, 38 (3): 419-424.)

[12] Boudchiche M, Mazroui A, Bebah M O A O, *et al.* AlKhalil Morpho Sys II: a robust Arabic morpho-syntactic analyzer [C]// Proc of International Conference on Information Technology for Organizations Development. 2016: 23-28.

[13] Zhang Xinwei, Wu Bin. Short text classification based on feature extension using the N-gram model [C]// Proc of International Conference on Fuzzy Systems and Knowledge Discovery. Piscataway, NJ: IEEE Press, 2016: 710-716.

[14] 刘德喜，聂建云，张晶，等．中文微博情感词提取：N-gram 为特征的分类方法 [J]．中文信息学报, 2016, 30 (4): 193-205. (Liu Dexi, Nie Jianyun, Zhang Jing, *et al.* Extracting sentimental lexicons from Chinese microblog: a classification method using N-gram features [J]. Journal of Chinese Information Processing, 2016, 30 (4): 193-205.)

[15] 何天文，王红．基于语义语法分析的中文语句困惑度评价 [J]．计算机应用研究, 2017, 34 (12): 3538-3542. (He Tianwen Wang Hong. Evaluating perplexity of Chinese sentences based on grammar & semantics analysis [J]. Application Research of Computers, 2017, 34 (12): 3538-3542.)